第 34 卷 第 15 期 计 算 机 工 程 2008 年 8 月
Vol.34 No.15 Computer Engineering August 2008
·软件技术与数据库· 文章编号：1000—3428(2008)15—0039—03 文献标识码：A 中图分类号：TP391

基于潜在语义分析的 Web 服务筛选技术

张孝国，黄广君，曹利红，郭洪涛
(河南科技大学电子信息工程学院，洛阳 471003)

摘 要：Web 服务匹配算法普遍缺少服务筛选机制，基于关键词对服务基本描述和服务质量描述进行匹配，会导致服务匹配效率低且准确率不高。该文提出一种基于潜在语义分析的 Web 服务筛选方法，将服务的基本描述和服务质量描述以树形结构属性模板表示，采用一定的词频统计和权重方法构建潜在语义空间，生成广告服务索引数据库，根据服务请求进行筛选。实验结果表明，该服务筛选方法具有较好的筛选准确率和筛选完全率，能够较大程度地提高服务匹配效率。

关键词：潜在语义分析；Web 服务；服务筛选；权重

Web Services Filtrate Technologies Based on Latent Semantic Analysis

ZHANG Xiao-guo, HUANG Guang-jun, CAO Li-hong, GUO Hong-tao
(Electronic Information Engineering College, Henan University of Science & Technology, Luoyang 471003)

【Abstract】The available services discovery mechanisms have lower matching efficiency and precision because of lacking services filtrate mechanism and matching the services basic attributes, quality of services attributes by keywords. This paper puts forward a Web services filtrate method based on Latent Semantic Analysis(LSA). This method describes Web services basic attributes, quality attributes using tree structure, uses certain terms-frequency statistic method and weights method to build latent semantic analysis space, and builts advertising services index database and filtrates the Web services according to services request. Experimental results prove that this algorithm has higher precision and recall and improves services matching efficiency largely.

【Key words】Latent Semantic Analysis(LSA); Web services; services filtrate; weight

1 概述

随着 Web 服务数量和种类的增加，如何方便高效地实现服务发现，是面向服务的体系架构所要解决的主要问题之一。目前大多数 Web 服务描述语言(如 WSDL, SCDL[1], OWL-S 等)都自发地遵循服务描述模型{*S*, *C*, *P*}，其中，*S* 是基本描述；*C* 是服务功能描述；*P* 是属性描述。但该模型对服务质量考虑不够，文献[2]将服务描述分为基本描述、基调描述和服务质量描述 3 部分。此外，还有不少研究者引入了服务质量本体对 OWL-S 规范进行了扩展[3]。

由此可见，目前 Web 服务描述模型包括服务基本描述和服务质量描述，而这些描述一般都是异构的文本信息。绝大多数服务匹配算法，都是同时对服务基本描述、服务质量描述、功能与行为描述进行匹配，缺乏一定的过滤筛选技术，服务匹配计算工作量大，匹配效率低，而且大多算法基于关键词计算服务基本描述和服务质量描述的相似度。因此，本文提出一种基于潜在语义分析的 Web 服务筛选技术，在服务匹配时，首先筛选出基本描述和服务质量描述满足需要的服务，再进一步进行服务匹配。

2 理论基础

2.1 潜在语义分析

潜在语义分析(Latent Semantic Analysis, LSA)是指：词语出现在某一个文档中以及 2 个词语出现在同一段上下文中不是完全随机的，而是存在某种潜在语义结构，体现了一种“词语-文档”双重概率关系。如果把这种潜在语义结构提取出来，建立词与词之间的语义关系，就可以消除词语用法的多样性和词语使用的随意性对检索产生的偏差[4]。

LSA 利用截断的奇异值分解(Truncated Singular Value Decomposition, TSVD)降秩方法实现信息抽取和噪声去除，将文档的高维表示投影在低维的潜在语义空间中，从而呈现出潜在的语义结构，通过多个维度的组合隐式地再现概念与概念间的差异和关联。

2.2 服务属性分析

潜在语义分析是通过构建文档集的“词语-文档”矩阵，实现文档与词语的语义联系。而 Web 服务的基本描述和服务质量描述都包含若干个孤立的文档，在服务筛选时并不能直接建立这些孤立文档的“词语-文档”矩阵，需要将它们作为一个整体综合考虑，建立“词汇-服务”矩阵。本文将 Web 服务的基本描述和服务质量描述用图 1 所示的“树形结构属性模板”表示为“服务属性树”，然后将“服务属性树”看作“服务文档”，构建“词汇-服务”矩阵并进行相应处理，从而将服务的筛选过程转化为对“服务属性树”的筛选过程。本文的“树形结构属性模板”可以根据描述信息的具体情况，灵活地增加或减少信息分支，同时可以较直观地体现信息的

基金项目：教育部科学技术基金资助重点项目(03081)
作者简介：张孝国(1980－)，男，助教、硕士研究生，主研方向：语义 Web，分布式计算及应用；黄广君，副教授、博士；曹利红，学士；郭洪涛，讲师、硕士研究生
收稿日期：2008-01-11 **E-mail：**zhxiaoguo@163.com

层次关系，该模板基本结构如图 1 所示。

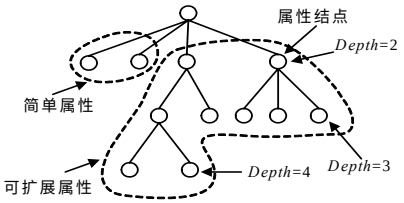

图 1　树形结构属性模板

简单属性指一个简单文档就可以表示的属性，如服务名称等；可扩展属性指需要分层表示的属性，有多个不同层次的简单文档构成。在这些简单文档中词汇大多只出现一次，而且在树形结构中，词汇的重要性一般通过结点层次体现出来，上层结点一般为服务的主要信息，重要性较大，下层结点一般为主要信息的补充说明，重要性较小。由此可见，直接利用原有词频并不能较好地表示词汇对服务文档的重要性，针对服务基本描述和服务质量描述的特殊性，本文在词频统计时分别采用了“原始词频”统计和“等价词频”统计 2 种方法，并通过实验对比了基于这 2 种方法的服务筛选性能。定义统计量如下：

(1) tf_{ij}：服务属性树 j 中词语 i 的原有频率；

(2) tf_{ikj}：服务属性树 j 中词语 i 在第 k 层属性结点中的原有频率；

(3) $tf_{ij}^{'}$：服务属性树 j 中词语 i 的等价频率；

(4) $tf_{ikj}^{'}$：服务属性树 j 中词语 i 在第 k 层属性结点中的等价频率；

(5) T_{depth}：服务属性树 j 的深度；

(6) $Nodeik_{depth}$：第 k 层词语 i 所在属性结点的深度。

原始词频统计：假设服务属性树的深度为 n，直接统计每层属性结点中词汇 i 的原有频率，然后相加即可。具体计算如下：

$$tf_{ij} = \sum_{k=1}^{n} tf_{ikj} \tag{1}$$

等价词频统计：在统计词频时，首先计算服务属性树中每层属性结点中词汇 i 的等价词频 $tf_{ikj}^{'}$，然后将所有层的等价词频相加，并对结果向上取整作为服务 j 中词汇 i 的等价词频 $tf_{ij}^{'}$。等价词频 $tf_{ikj}^{'}$ 的具体计算如下：

$$tf_{ikj}^{'} = tf_{ikj} \times (1 + \mathrm{lb}[T_{depth} - (Nodeik_{depth} - 1)]) \tag{2}$$

等价词频 $tf_{ij}^{'}$ 的具体计算如下：

$$tf_{ij}^{'} = \left\lceil \sum_{k=1}^{n} tf_{ikj}^{'} \right\rceil \tag{3}$$

本文针对这 2 种词频统计方法，分别构建了“词汇-服务”矩阵，设计了基于原始词频和等价词频的服务筛选方法并进行了实验对比，除词频统计方法不同之外，2 种筛选方法的其他步骤基本相同。

3　Web 服务筛选

3.1　Web 服务筛选流程

本文的服务筛选流程如图 2 所示。收集训练服务集构建潜在语义空间；将广告服务的基本描述和服务质量描述映射到潜在语义空间生成广告索引数据库；同时按要求处理服务请求生成请求索引向量；最后计算请求索引向量与广告索引向量的相似度，将相似度大于等于阈值的服务按照相似度由大到小插入到结果列表，直到广告服务全部筛选完毕，返回结果列表。

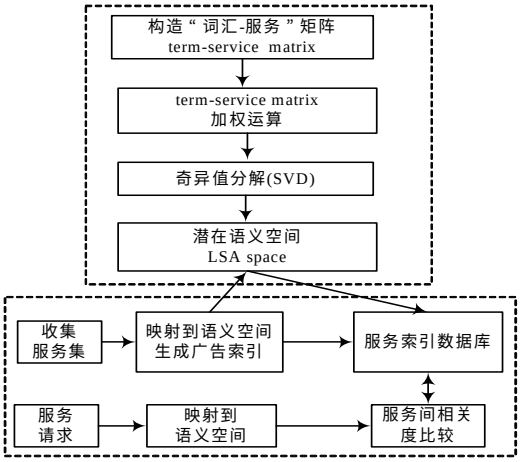

图 2　基于 LSA 的服务筛选流程

3.2　Web 服务筛选实现步骤

3.2.1　构建潜在语义空间

收集 Web 服务实例，从服务基本描述和服务质量描述信息中提取能够表征相应服务属性的词汇，构建“词汇-服务”矩阵。该矩阵格式如下：

$$\boldsymbol{X}_{m\times n}=[X_{ij}]=(\text{service1, service2,}\cdots\text{, service}N) = (\text{term1, term2,}\cdots\text{, term}M)^{\mathrm{T}}$$

其中，X_{ij} 的值为词汇 i 在服务 j 中的词频。

本文采用“原始词频”和“等价词频”2 种统计方法计算词频；同时还统计出了词汇 i 出现的服务数、服务集中词汇 i 出现的总次数以及服务 j 中有效词汇的个数，以便对“词汇-服务”矩阵进行加权变换，使矩阵中的元素更接近于自然语言中词汇与服务间的关系。采用如下的加权函数进行矩阵加权变换：

$$W(i,j) = LW(i,j) \times GWT(i) \times GWD(j) \tag{4}$$

其中，$LW(i,j)$ 为局部权重；$GWT(i)$ 为词语全局权重；$GWD(j)$ 为服务文档全局权重。

先定义以下统计量：

(1)gf_i：在整个“服务属性树”集中词汇 i 出现频数之和；

(2)sgf：在“服务属性树”集中所有词汇出现频数之和；

(3)dl_j：服务属性树 j 的长度，即服务属性树 j 包含的词汇总数。

本文采用对数词频作为局部权重，削减高频词对同一服务中其他词汇语义贡献的掩盖作用，即：

$$LW(i,j) = \mathrm{lb}(tf_{ij} + 1) \tag{5}$$

一般认为，某一词汇提供给服务文档集的信息量越大，说明它分辨服务文档的能力越强，它的全局权重应当越高，因此，本文借鉴信息论中“熵”及其相关概念的特殊性质，采用下式计算词语全局权重：

$$GWT(i) = 1 + \frac{\sum_j P(docj \mid termi)\mathrm{lb}P(docj \mid termi)}{\mathrm{lb}n} \tag{6}$$

其中，$P(docj \mid termi) = tf_{ij} / gf_i$ 是条件“词汇 i 出现”成立的情况下，“文档 j 出现”的概率。同理，可采用“信息增益”定义服务文档全局权重，即：

$$GWD(j)=1-\frac{\sum_{i=1}^{m}\frac{gf_i}{sgf}\times \mathrm{lb}\frac{gf_i}{sgf}}{\sum_i \frac{tf_{ij}}{dl_j}\mathrm{lb}\frac{tf_{ij}}{dl_j}} \tag{7}$$

在矩阵加权转换之后，本文使用 Matlab 进行 TSVD 变换，然后进行 SVD 反运算，得到原始矩阵的一个近似阵 $\hat{\boldsymbol{X}}_k = T_k S_k D_k^{\mathrm{T}}$，其中，$T_k=(t_1, t_2,\cdots, t_k)$；$D_k=(d_1, d_2,\cdots,d_k)$。从而得到由 $span\{t_1, t_2,\cdots, t_k\}$和 $span\{d_1, d_2,\cdots, d_k\}$构成的潜在语义空间。

3.2.2　生成广告索引数据库

广告索引数据库主要包括 2 部分：(1)已参加过 TSVD 变换的服务，可把 D_k 中的行视为代表文档的向量，即将 $\boldsymbol{doc}'_i=(\sigma_1 d_{1,i}, \sigma_2 d_{2,i},\cdots, \sigma_k d_{k,i})^{\mathrm{T}}$ 视为服务 i 在 k 维向量空间中的表示；(2)新收集的广告服务，采用式(8)将其服务文档 doc 映射到潜在语义空间生成广告索引向量 $\boldsymbol{doc}^*$。

$$\boldsymbol{doc}^* = \boldsymbol{doc}^{\mathrm{T}}\boldsymbol{W}_t\boldsymbol{T}_k \tag{8}$$

其中，$\boldsymbol{W}_t$ 为对角矩阵，对角线上元素对应各词汇的全局权重。

3.2.3　服务相关度比较

本文采用余弦相关度计算请求服务与广告服务的相似度，筛选出相似度大于等于阈值的广告服务，按相似度由大到小插入到结果列表中。设 $\boldsymbol{Ser}_i^*=(Ser_{i,1}^*, Ser_{i,1}^*,\cdots, Ser_{i,k}^*)^{\mathrm{T}}$ 和 $\boldsymbol{Ser}_j^*=(Ser_{j,1}^*, Ser_{j,1}^*,\cdots, Ser_{j,k}^*)^{\mathrm{T}}$ 为 2 个服务文档的 LSA 低维向量表示，则本文服务相关度计算公式为

$$Sim(\boldsymbol{Ser}_i^*, \boldsymbol{Ser}_j^*)=\frac{\sum_{h=1}^{k} Ser_{i,h}^*\times Ser_{j,h}^*}{\sqrt{\sum_{h=1}^{k}(Ser_{i,h}^*)^2}\times\sqrt{\sum_{h=1}^{k}(Ser_{j,h}^*)^2}} \tag{9}$$

4　仿真性能测试

本文借鉴信息检索中查准率和查全率的定义思想[5]，定义服务筛选准确率 P_{filtrate} 和完全率 R_{filtrate} 如下：

(1)P_{filtrate}：应答集合中相关服务数与应答集合中服务数的比值。

(2)R_{filtrate}：应答集合中相关服务数与总的相关服务数的比值。

本文分别设计了基于原始词频和等价词频统计的 Web 服务筛选原型系统 $\mathrm{WSFS}_{\mathrm{primal}}$ 和 $\mathrm{WSFS}_{\mathrm{equal}}$。参考 OWLS-TC V2 服务测试数据集，增加了服务质量描述，制定了测试集 WSFS-TC，基于该测试集构建了潜在语义空间，对本文系统进行了仿真性能测试，同时与基于关键词组合的筛选算法以及 StarWSDS 系统[2]针对服务基本描述和服务质量描述信息的匹配性能进行了比较，统计结果如表 1 所示。

表 1　仿真性能测试结果

筛选算法	P_{filtrate}/(%)	R_{filtrate}/(%)
关键词组合	32	91
StarWSDS	65	93
$\mathrm{WSFS}_{\mathrm{primal}}$	71	94
$\mathrm{WSFS}_{\mathrm{equal}}$	76	95

由此可见，本文提出的基于潜在语义分析的服务筛选算法具有较高的筛选准确率和完全率。

5　结束语

本文分析了当前的 Web 服务描述模型和大多数的服务匹配算法，提出一种基于潜在语义分析的 Web 服务筛选方法，开发了原型系统，进行了仿真性能测试。实验结果表明该服务筛选方法能够过滤掉绝大多数不相关服务，缩小服务匹配范围，节省服务匹配时间，提高了服务匹配效率。

参考文献

[1] Gao Xiang, Yang Jian, Papazoglou M P. The Capability Matching of Web Services[C]//Proc. of the 4th International Symposium on Multimedia Software Engineering. California, USA: [s. n.], 2002.

[2] 胡建强, 邹　鹏, 王怀民, 等. Web 服务描述语言 QWSDL 和服务匹配模型研究[J]. 计算机学报, 2005, 28(4): 505-513.

[3] 员红娟, 叶飞跃, 李　霞, 等. 基于语义的服务发现核心技术研究[J]. 计算机应用, 2006, 26(11): 2661-2662.

[4] 余正涛, 樊孝忠, 郭剑毅, 等. 基于潜在语义分析的汉语问答系统答案提取[J]. 计算机学报, 2006, 29(10): 1889-1891.

[5] 吴　健, 吴朝晖, 李　莹, 等. 基于本体论和词汇语义相似度的 Web 服务发现[J]. 计算机学报, 2005, 28(4): 595-601.

（上接第 32 页）

计算个体 S 和 R_i 的距离，将该距离与欧几里得距离比较，得到表 6。可以看出，该算法得到的软件人性格比较距离与对应欧氏距离一致。

表 6　软件人性格曲线的距离

距离	$D(S, R_i)$		
	R_1	R_2	R_3
逼近距离	0.668 2	0.334 9	1.000 0
欧氏距离	1.0	0.5	2.0

6　结果分析

以上运算适用于三维、五维向量及任意可数维。具有相似性格曲线的软件人性格总是距离较小，反映为表 2 中的 $\delta_{\max}(\mathrm{in})=0.309\ 8$，而根据表 3~表 5，曲线相异的软件人之间 $\delta_{\min}(\mathrm{out})=0.592\ 3$，在 $[\delta_{\max}(\mathrm{in}), \delta_{\min}(\mathrm{out})]$ 之间设定阈值即可区分两者，说明算法对软件人的性格具有良好的可分性。在 5.2 节中，性格向量被扩大为五维。通过计算性格模式 S 和 R_1, R_2, R_3 的距离，可以看出该算法表示的距离与欧氏距离一致，说明对于性格曲线的分辨能力与欧氏距离相当。

7　结束语

软件人性格由环境赋予，可以用向量表示，是可列无限维的。通过在赋范线性空间的切比雪夫逼近，对软件人的典型性格曲线进行距离度量。实验证明，该算法符合常规的距离运算规律，有助于解决软件人在网上的数据无限性和性格变化性问题。

参考文献

[1] 曾广平, 涂序彦. 软件人[M]//中国人工智能进展. 北京: 北京邮电大学出版社, 2003: 677-682.

[2] 王志良, 赵彦玲, 郝春辉, 等. 采用人工心理理论的商品选购专家系统[J]. 北京科技大学学报, 2001, 23(4): 376-377.

[3] 段　军, 戴居丰, 涂序彦. 软件人的体系结构及其网络平台[J]. 天津大学学报, 2006, 39(S1): 248-251.

[4] 赵积春, 王志良, 王　超. 情绪建模与情感虚拟人研究[J]. 计算机工程, 2007, 33(1): 212-215.

[5] Cheney E W. 逼近论导引[M]. 徐献喻, 译. 上海: 上海科学技术出版社, 1981.